Generalized Bayesian predictive density operators

Fuyuhiko Tanaka[*]

Department of Mathematical Informatics, University of Tokyo,

7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan

(Dated: February 1, 2008)

Abstract

Recently the quantum Bayesian prediction problem was formulated by Tanaka and Komaki (2005). It is shown that Bayesian predictive density operators are the best predictive density operators when we evaluate them by using the averaged quantum relative entropy based on a prior distribution. In the present paper, we adopt the quantum α-divergence as a wider class of loss function. The generalized Bayesian predictive density operator is defined and shown to be best among all the estimates of the unknown density operator.

---
[*]Electronic address: ftanaka@stat.t.u-tokyo.ac.jp

I. INTRODUCTION

In classical statistics, the problem of predicting an unobserved variable y by using an observed variable x has been investigated. Suppose that a parametric model

$$\mathcal{P} = \{p(y|\theta) : \theta \in \Theta\},$$

which is a set of probability densities, is given, where Θ is a parameter space. Random variables x and y are distributed according to the same true probability density $p(\cdot|\theta)$ in $\mathcal{P}$. We predict the unobserved variable y with a predictive density $\hat{p}(y; x)$ constructed by using the observed variable x. The closeness of the true density $p(y|\theta)$ and a predicted density $\hat{p}(y; x)$ is evaluated by using the Kullback-Leibler divergence

$$D(p||\hat{p}) := \int p(y|\theta) \log \frac{p(y|\theta)}{\hat{p}(y; x)} \mathrm{d}y.$$

Aitchison [1] showed that a Bayesian predictive density

$$p_\pi(y|x) := \int_\Theta p(y|\theta)\pi(\theta|x)\mathrm{d}\theta, \tag{1}$$

where $\pi(\theta|x)$ is a posterior distribution, is the best predictive density when we evaluate a predictive density $\hat{p}(y; x)$ by using the average Kullback-Leibler divergence $\int \pi(\theta) \int D(p||\hat{p})p(x|\theta)\mathrm{d}x\mathrm{d}\theta$, where $\pi(\theta)$ is a probability density. This result was extended to the quantum setting by Tanaka and Komaki [2].

Let us consider this result more deeply. From an observation x, we obtain the corrected information $\pi(\theta|x)$ on the unknown parameter θ and Eq.(1) is obtained by taking mixture of possible probability densities $p(y|\theta)$ with respect to $\pi(\theta|x)$. However, there are many ways of taking mixture. For example, let $p_1(x)$ and $p_2(x)$ denote two possible Gaussian distributions. Then, $p(x) = \frac{1}{2}p_1(x) + \frac{1}{2}p_2(x)$ is one possibility. Another possibility of mixture is given by $\log p'(x) = \frac{1}{2}\log p_1(x) + \frac{1}{2}\log p_2(x)$. The advantage of the latter is that the mixture itself is again a Gaussian distribution. In classical statistics, such a mixture is often useful and a general class of mixture, an α-mixture [3], can be defined by

$$
\begin{aligned}
p_\pi^{(\alpha)}(y|x) &:= \left\{\int \{p(y|\theta)\}^{\frac{1-\alpha}{2}}\pi(\theta|x)\mathrm{d}\theta\right\}^{\frac{2}{1-\alpha}}, \\
p_\pi^{(+1)}(y|x) &:= \exp\left\{\int \log(p(y|\theta))\pi(\theta|x)\mathrm{d}\theta\right\}.
\end{aligned}
\tag{2}
$$

Corcuera and Giummolè [4] showed that the above predictive distribution (2), which they called a generalized Bayesian predictive density, is optimal under the following α-divergence:

$$D^{(\alpha)}(p||q) := \frac{4}{1-\alpha^2}\left\{1 - \int p(x)^{\frac{1-\alpha}{2}}q(x)^{\frac{1+\alpha}{2}}\mathrm{d}x\right\},$$

where $\alpha \neq \pm 1$. When $\alpha = \pm 1$, it is defined by

$$
\begin{aligned}
D^{(-1)}(p||q) &:= \int p(x) \log\left(p(x)/q(x)\right) \mathrm{d}x, \\
D^{(+1)}(p||q) &:= D^{(-1)}(q||p).
\end{aligned}
\tag{3}
$$

The α-divergence is closely related to the α-entropy of Rényi [5] and the Chernoff distance [6] in the classical information theory. Our purpose in the present paper is to extend the result obtained by Corcuera and Giummolè to the quatum setting.

In quantum statistics, which was initiated by Helstrom, Holevo, and other researchers [7, 8] a quarter century ago, the optimal estimation of the parameter of the unknown quantum state has been one of the hot topics over the past several years with recent developments of experimental techniques [9, 10, 11]. They usually consider the ideal situation with all measurements (described by POVM) allowed and often deal with large sample cases. The theoretical limitation on the accuracy of the parameter estimation has been clarified to some extent. However, in practical situation, we often need to know the density operator describing the unknown quantum state rather than the unknown parameter with a given measurement device. Bužek *et al.* [12] recommended to use Bayesian technique especially when the sample size of experimental data is small. They proposed to use a posterior state corresponding to a posterior distribution in classical counterparts. Tanaka and Komaki [2] formulated the estimation of the unknown density operator as the quantum prediction problem and showed that the posterior state is best among all the estimates of the unknown density operator.

However, the optimality argument depends on the choice of an evaluation function, which is called *a loss function* in mathematical statistics. Here, we adopt a general class of loss fucntion as a quantum counterpart of the classical α-divergence (3). Then, we define the generalized Bayesian predictive density operator and show that it is the optimal density estimate. Our result includes the previous result obtained by Tanaka and Komaki [2] as a special case.

In the next section, we briefly review our setting, essentially, the same one as in Tanaka and Komaki [2], except for the choice of a loss function. In Section 3, we prove our main result. Concluding remarks are described in Section 4.

3

II. PRELIMINARY

We briefly summarize some notations of quantum measurement. Let $\mathcal{H}$ be a separable (possibly infinite dimensional) Hilbert space of a quantum system. An Hermitian operator ρ on $\mathcal{H}$ is called a *state* or *density operator* if it satisfies,

$$\mathrm{Tr}\rho = 1, \quad \rho \geq 0.$$

We denote the set of all states on $\mathcal{H}$ as $\mathcal{S}(\mathcal{H})$.

Let Ω be a space of all possible outcomes of an experiment (e.g., $\Omega = \mathbf{R}^n$) and suppose that a σ-algebra $\mathcal{B} := \mathcal{B}(\Omega)$ of subsets of Ω is given. An affine map μ from $\mathcal{S}(\mathcal{H})$ into a set of probability distributions on Ω, $\mathcal{P} = \{\mu(\mathrm{d}x)\}$ is called a *measurement*. There is a one-to-one correspondence between a measurement and a resolution of the identity [8]. A map from $\mathcal{B}$ into the set of positive Hermitian operators

$$E : B \mapsto E(B),$$

where E satisfies

$$E(\phi) = O, E(\Omega) = I, \tag{4}$$

$$E(\cup_i B_i) = \sum_i E(B_i), \quad B_i \cap B_j = \phi, \quad \forall B_i \in \mathcal{B}, \tag{5}$$

is called a *positive operator valued measure* (*POVM*). Any physical measurement can be represented by a POVM.

Now we describe our setting of state estimation. Assume that a state ρ_θ on $\mathcal{H}$ is characterized by an unknown finite-dimensional parameter $\theta \in \Theta \subset \mathbf{R}^n$. If $\dim \mathcal{H} < \infty$, θ may cover full range (often called the full model.).

A quantum state for N systems, $\rho^{(N)}$, is described on the N-fold tensor product Hilbert space $\mathcal{H}^{\otimes N}$. Suppose that a system composed of $N+M$ subsystems is given and that a measurement is performed only for selected N subsystems with the other M subsystems left. Then, the measurement is described by $\{E_x \otimes I\}$, where $\{E_x\}$ is a POVM on $\mathcal{H}^{\otimes N}$ and I is the identity operator on $\mathcal{H}^{\otimes M}$.

Our aim is to estimate the true state $\sigma_\theta := \rho_\theta^{\otimes M}$ of the remaining M subsystems by using a measurement $\{E_x\}$ on the selected N subsystems $\rho_\theta^{\otimes N}$. We fix an arbitrarily chosen measurement. Note that the above measurement is not necessarily in the form of a tensor

4

product $E_x^{\otimes N}$, which represents a repetition of the same measurement E_x for each system. Thus, all possible measurements on N subsystems, which may use entanglement, are considered. We call $\hat{\sigma}(x)$ an estimate of the true state as *a predictive density operator*. The problem of the quantum prediction is to seek for the optimal predictive density operator.

The performance of a predictive density operator $\hat{\sigma}(x)$ is evaluated by the quantum α-divergence $D^{(\alpha)}(\sigma_\theta||\hat{\sigma}(x))$, a quantum analogue of the α-divergence (3) in classical statistics. The quantum α-divergence from ρ to σ is defined by

$$D^{(\alpha)}(\rho||\sigma) := \frac{4}{1-\alpha^2} \left(1 - \mathrm{Tr}\sigma^{\frac{1+\alpha}{2}} \rho^{\frac{1-\alpha}{2}} \right), \quad \text{if } \alpha \neq \pm 1. \tag{6}$$

When $\alpha = \pm 1$, it is defined by

$$D^{(\alpha=-1)}(\rho||\sigma) := \mathrm{Tr}\rho(\log\rho - \log\sigma) =: D^{(\alpha=1)}(\sigma||\rho).$$

It satisfies the positivity condition $D^{(\alpha)}(\rho||\sigma) \geq 0$ and $D^{(\alpha)}(\rho||\sigma) = 0 \Leftrightarrow \rho = \sigma$. Other properties and useful inequalities, see, e.g., [13]. Thus, it can be used as a measure for the goodness of a predictive density operator. Note that the quantum α-divergence is reduced to the classical α-divergence when the two density operators are commutative.

Remark. 1.

The quantum α-divergence can be given a suitable meaning as a measure only when $|\alpha| \leq 3$. (See, Hasegawa [14].) However, our statement formally holds for any α. We also assume additional conditions on density operators such that the α-divergence is finite.

Remark. 2.

The quantum α-divergence can be rewritten in the relative g-entropies $H_g(\rho||\sigma) := \mathrm{Tr}\rho^{\frac{1}{2}}g(L_\sigma/R_\rho)(\rho^{\frac{1}{2}})$, where g in an operator convex function and $g(1) = 0$ and $L_\sigma(X) = \sigma X, R_\rho(X) = X\rho$ are superoperators. The relative g-entropy was introduced by Petz [15].

If we assume a prior probability density $\pi(\theta)$ on the parameter space Θ, the mixture state for the whole N systems is given by

$$\rho^{(N)} := \int \mathrm{d}\theta \ \pi(\theta) \ \rho_\theta^{\otimes N}. \tag{7}$$

A state of the form (7) is called an *exchangeable state* [16], and arises, e.g., if each subsystem is prepared in the same unknown way, as in quantum state tomography. In a quantum

exchangeable model (7), as Schack *et al.* [16] showed, a posterior distribution $\pi(\theta|x)$ naturally arises. We assume that the whole system is in the exchangeable state and in our setting $\pi(\theta|x)$ is given by

$$\pi(\theta|x) := \frac{p(x|\theta)\pi(\theta)}{\int \mathrm{d}\theta \ p(x|\theta)\pi(\theta)},$$

where $p(x|\theta) = \mathrm{Tr}\rho_\theta^{\otimes N} E_x$.

Finally, let us define generalized Bayesian predictive density operators. First of all, we consider an α-mixture of σ_θ with respect to a posterior density $\pi(\theta|x)$.

$$\sigma_\pi^{(\alpha)}(x) := \begin{cases} \{\int \sigma_\theta^{\frac{1-\alpha}{2}} \pi(\theta|x)\mathrm{d}\theta\}^{\frac{2}{1-\alpha}}, & \alpha \neq 1, \\ \exp\{\int \log(\sigma_\theta)\pi(\theta|x)\mathrm{d}\theta\}, & \alpha = 1. \end{cases}$$

Clearly the above mixture is a positive operator and $\mathrm{Tr}\sigma_\pi^{(\alpha)}(x) > 0$. Thus, we define the generalized Bayesian predictive density operator in the following normalized form.

$$\tilde{\sigma}_\pi^{(\alpha)}(x) := \frac{1}{C_\alpha(x)}\sigma_\pi^{(\alpha)}(x), \quad C_\alpha(x) := \mathrm{Tr}\sigma_\pi^{(\alpha)}(x).$$

In the following section, we show that the generalized Bayesian predictive density operator is the best predictive density operator in the sense that it minimizes the averaged quantum α-divergence from the true density operator.

III.   MAIN THEOREM

In classical statistics, Corcuera and Giummolè [4] showed that the generalized Bayesian predictive density $p_\pi^{(\alpha)}(y|x)$ is the best predictive density under the α-divergence when a proper prior $\pi(\theta)$ is given. We derive the corresponding result for quantum predictive density operators.

Theorem.

Let $\alpha \in \mathbf{R}$ be fixed. Suppose that we perform a measurement for selected N subsystems $\rho_\theta^{\otimes N}$ of a system $\rho_\theta^{\otimes(N+M)}$ composed of $N + M$ subsystems in order to estimate the remaining M subsystems $\sigma_\theta = \rho_\theta^{\otimes M}$. The true parameter value θ is unknown and a prior probability density $\pi(\theta)$ is assumed. Let $\hat{\sigma}(x)$ be any predictive density operator, where x is an outcome

of a measurement $\{E_x\}$ for the N subsystems. Performance of a predictive density operator $\hat{\sigma}(x)$ is measured with the averaged quantum α-divergence

$$\mathrm{E}_\theta \mathrm{E}_x[D^{(\alpha)}(\sigma_\theta||\hat{\sigma})] = \int \mathrm{d}\theta\ \pi(\theta) \int \mathrm{d}x\ p(x|\theta) D^{(\alpha)}(\sigma_\theta||\hat{\sigma}(x))$$

from the true state σ_θ. Then, the generalized Bayesian predictive density operator $\tilde{\sigma}_\pi^{(\alpha)}(x)$ is the best predictive density operator.

Proof.

When $\alpha \neq \pm 1$, let us consider taking an average of the difference between $D^{(\alpha)}(\sigma_\theta||\hat{\sigma}) - D^{(\alpha)}(\sigma_\theta||\tilde{\sigma}_\pi^{(\alpha)})$. From now on, we omit α in the $\tilde{\sigma}_\pi^{(\alpha)}$.

$$\begin{aligned}
&E^\pi E^{M_x}[D^{(\alpha)}(\sigma_\theta||\hat{\sigma}) - D^{(\alpha)}(\sigma_\theta||\tilde{\sigma}_\pi)]\\
&= \int \mathrm{d}\theta\pi(\theta) \int \mathrm{d}xp(x|\theta)\left\{\frac{4}{1-\alpha^2}\mathrm{Tr}\sigma_\theta^{\frac{1-\alpha}{2}}(\tilde{\sigma}_\pi^{\frac{1+\alpha}{2}} - \hat{\sigma}^{\frac{1+\alpha}{2}})\right\}\\
&= \int \mathrm{d}xp_x \int \mathrm{d}\theta\frac{\pi(\theta)p(x|\theta)}{p_x}\left\{\frac{4}{1-\alpha^2}\mathrm{Tr}\sigma_\theta^{\frac{1-\alpha}{2}}(\tilde{\sigma}_\pi^{\frac{1+\alpha}{2}} - \hat{\sigma}^{\frac{1+\alpha}{2}})\right\}\\
&= \int \mathrm{d}xp_x \int \mathrm{d}\theta\ \pi(\theta|x)\left\{\frac{4}{1-\alpha^2}\mathrm{Tr}\sigma_\theta^{\frac{1-\alpha}{2}}(\tilde{\sigma}_\pi^{\frac{1+\alpha}{2}} - \hat{\sigma}^{\frac{1+\alpha}{2}})\right\}\\
&= \int \mathrm{d}xp_x\frac{4}{1-\alpha^2}\mathrm{Tr}\left\{\left(\int \mathrm{d}\theta\pi(\theta|x)\sigma_\theta^{\frac{1-\alpha}{2}}\right)(\tilde{\sigma}_\pi^{\frac{1+\alpha}{2}} - \hat{\sigma}^{\frac{1+\alpha}{2}})\right\}\\
&= \int \mathrm{d}xp_x\frac{4}{1-\alpha^2}\mathrm{Tr}\left\{C_\alpha^{\frac{1-\alpha}{2}}\tilde{\sigma}_\pi^{\frac{1-\alpha}{2}}(\tilde{\sigma}_\pi^{\frac{1+\alpha}{2}} - \hat{\sigma}^{\frac{1+\alpha}{2}})\right\}\\
&= \int \mathrm{d}xp_xC_\alpha^{\frac{1-\alpha}{2}}\frac{4}{1-\alpha^2}\left\{1 - \mathrm{Tr}\tilde{\sigma}_\pi^{\frac{1-\alpha}{2}}\hat{\sigma}^{\frac{1+\alpha}{2}})\right\}\\
&= \int \mathrm{d}xp_xC_\alpha^{\frac{1-\alpha}{2}}D^{(\alpha)}(\tilde{\sigma}_\pi||\hat{\sigma}) \geq 0,
\end{aligned}$$

where $p_x := \int \mathrm{d}\theta'\pi(\theta')p(x|\theta')$ is the marginal density of x. The last inequality holds due to the positivity of the quantum α-divergence $D^{(\alpha)}(\sigma||\sigma') \geq 0$ and $p_x \geq 0$. Since $\hat{\sigma}(x)$ is arbitrarily chosen, it is shown that $\tilde{\sigma}_\pi^{(\alpha)}(x)$ is better than any other $\hat{\sigma}(x)$. We can repeat the same procedure for $\alpha = \pm 1$.

Q.E.D.

IV.   REMARKS

When $\alpha = 0$, the quantum α-divergence is closely related to the fidelity $F(\rho, \sigma) :=$ $\mathrm{Tr}|\sqrt{\rho}\sqrt{\sigma}|$, where $|A| := \sqrt{A^*A}$. Since $|\mathrm{Tr}A| \leq \mathrm{Tr}|A|$, we obtain

$$D^{(0)}(\rho||\sigma) = 4(1 - \mathrm{Tr}\sqrt{\rho}\sqrt{\sigma}) \leq 4(1 - F(\rho, \sigma)).$$

The equality holds when ρ and σ are commutative or both ρ and σ are pure states. The fidelity is often used as a measure in the quantum information theory [17]. How our theorem can be extended when we adopt the fidelity as a loss function is left for the future study.

Acknowledgments

[1] J. Aitchison, Biometrica **62**, 547 (1975).

[2] F. Tanaka and F. Komaki, Phys. Rev. A **71**, 052323 (2005).

[3] S. Amari, Proceedings of 2nd International Symposium on Information Geometry and its Applications (IGAIA) (2005).

[4] J. M. Corcuera and F. Giummolè, Scand. J. Statist. **26**, 265 (1999).

[5] A. Rényi, Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability **1**, 547 (1961).

[6] H. Chernoff, Ann. Math. Stat. **23**, 493 (1952).

[7] C. W. Helstrom, *Quantum Detection Theory* (Academic Press, New York, 1976).

[8] S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam, 1982).

[9] M. Hayashi, *Asymptotic Theory of Quantum Statistical Inference* (World Scientific, Singapore, 2005).

[10] O. E. Barndorff-Nielsen, R. D. Gill, and P. E. Jupp, J. Roy. Statist. Soc. Ser. B **65**, 775 (2003).

[11] M. Paris and J. Řeháček, *Quantum State Estimation* (Springer, Berlin, 2004).

[12] V. Bužek, R. Derka, G. Adam, and P. L. Knight, Ann. Phys. (N.Y.) **266**, 454 (1998).

[13] S. Amari and H. Nagaoka, *Methods of Information Geometry* (AMS, Oxford, 2000).

[14] H. Hasegawa, Rep. Math. Phys. **33**, 87 (1993).

[15] D. Petz, Rep. Math. Phys. **23**, 57 (1986).

[16] R. Schack, T. A. Brun, and C. M. Caves, Phys. Rev. A **64**, 014305 (2001).

[17] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000).